Joint Registration and Representation Learning for Unconstrained Face Identification

Munawar Hayat*†, Salman H. Khan*‡, Naoufel Werghi††, Roland Goecke†
†University of Canberra, Australia, ‡Data61 - CSIRO and ANU, Australia
††Khalifa University, Abu Dhabi, United Arab Emirates
{munawar.hayat,roland.goecke}@canberra.edu.au, salman.khan@csiro.au, naoufel.werghi@kustar.ac.ae

Abstract

Recent advances in deep learning have resulted in human-level performances on popular unconstrained face datasets including Labeled Faces in the Wild and YouTube Faces. To further advance research, IJB-A benchmark was recently introduced with more challenges especially in the form of extreme head poses. Registration of such faces is quite demanding and often requires laborious procedures like facial landmark localization. In this paper, we propose a Convolutional Neural Networks based data-driven approach which learns to simultaneously register and represent faces. We validate the proposed scheme on template based unconstrained face identification. Here, a template contains multiple media in the form of images and video frames. Unlike existing methods which synthesize all template media information at feature level, we propose to keep the template media intact. Instead, we represent gallery templates by their trained one-vs-rest discriminative models and then employ a Bayesian strategy which optimally fuses decisions of all medias in a query template. We demonstrate the efficacy of the proposed scheme on IJB-A, YouTube Celebrities and COX datasets where our approach achieves significant relative performance boosts of 3.6%, 21.6% and 12.8% respectively.

1. Introduction

Owing to its wide range of potential applications, face recognition has been rigorously researched in computer vision community. Challenges in face recognition are associated with commonly occurring nuisances of facial data which include head pose rotations, illumination variations and expression deformations. In its initial days, facial data was systematically captured in controlled environments and algorithms were developed to individually tackle each of these nuisances [24]. Such algorithms could achieve impressive performance in constrained environments but failed in real-life scenarios. To advance research in unconstrained face recognition, Labelled Faces in the Wild (LFW) [15] and YouTube Faces (YTF) [39] datasets were released in 2007 and 2011 respectively. At the time of their release, the existing methods (developed using constrained data) performed poorly on LFW and YTF. Since then, a large focus of face recognition research has been on the development of algorithms which achieve superior performance on LFW and YTF. With the recent advances in deep learning, the current state of the art algorithms [33, 27] can now achieve human level performance on these datasets. Unconstrained face recognition is however still considered largely unresolved [22]. This is mainly because both LFW and YTF have a well-know frontal selection bias. Specifically, face images in both of these datasets were automatically detected using Viola and Jones [34], which frequently fails for non-frontal faces. The state of the art on YTF and LFW therefore performs poorly in the presence of large head rotations and extreme head poses [22, 6].

In this paper, we aim to address face recognition across extreme head rotations. Registration of such facial images is quite a challenging task and often requires sophisticated pre-processing steps such as landmark localization and frontalization. We propose to automatically learn facial image registration along with feature encoding as part of an end-to-end trainable Convolutional Neural Network. The proposed network (Sec. 3) has two modules: a registration module to learn a set of transformation parameters, and a representation module to learn meaningful feature encoding of input face images. The network is trained on 2.6 million images of 2622 subjects [27]. The proposed scheme is then evaluated on IJB-A [22], YouTube Celebrities [20] and COX [16] datasets for template based face identification. The IJB-A benchmark is specifically quite challenging and contains face images and video frames across extreme head poses and profile views (see Fig. 4). The proposed method achieves a significant performance boost on all of the evaluated datasets (Sec. 5).

*Equal contribution

The problem of face recognition is studied under verification and identification tasks. For verification, we compute a one-one similarity of a given probe face to verify its claimed identity. For identification, one-to-many similarities of the probe are computed in order to find its best match within a gallery of enrolled subjects. Face identification is therefore more challenging compared with face verification. Unconstrained face identification has however been largely neglected over the past few years. This is mainly because most of the research was driven by LFW and YTF datasets and their evaluation protocols are defined for verification only. In this paper, we address template based unconstrained face identification. A template may contain multiple heterogeneous medias in the form of still images or video frames. Face identification from templates is relevant in many commercial systems (*e.g*. FBI's most wanted list) where multiple images of an individual are simultaneously available. Although a template contains more information, it simultaneously poses challenges to effectively utilize this information. Unlike existing methods which merge all template media at feature level, we propose to keep it intact. To leverage from this myriad of information, we train one-vs-rest discriminative models for gallery templates (Sec. 4.3) and employ a Bayesian approach which optimally fuses classification decisions for medias of a given query template (Sec. 4.4).

2. Related Work

A generic face recognition system has three major components: **i)** registration of raw facial images, **ii)** feature encoding of the registered faces, and finally **iii)** classification (verification or identification). In the existing literature, techniques have been developed to individually deal with each of these three components. For **registration**, 2D and 3D face alignment methods have been devised [27, 33, 1]. These methods usually warp automatically detected facial landmarks onto a model face which has a canonical frontal view. For facial feature **representation**, the descriptors can either be manually designed or automatically learnt from large scale facial data. Local Binary Patterns [25], Histogram of Oriented Gradients [7] and Gabor wavelets [42] are some popular examples of the designed features. Most of the recent top performing face recognition methods employ features learnt from a large amount of training data using a Convolutional Neural Network (CNN). Examples include DeepFace [33], VGG-Face [27], FaceNet [30] and DeepID [32]. DeepFace and VGG-Face are based on common CNN architectures whereas FaceNet and DeepID use a specialized inception architecture. As a final step in feature learning, some of these methods employ metric learning (*e.g*. triplet loss embedding [29]) to learn optimal task specific feature embedding (*e.g*. for face verification using LFW and YTF datasets [33, 27]). After registration and feature encoding, the final step is **classification**. Any off-the-shelf classifier can be adapted for verification or identification. Different from previous works, this paper combines registration and representation steps. We propose to learn these as part of a single network. This avoids preprocessing procedures such as landmark localization which are not only computationally expensive but can also introduce many challenges specially in scenarios with extreme head poses (*e.g*. in IJB-A dataset).

With advancements in deep learning for image classification [23, 18, 13], face recognition performances on YTF and LFW datasets have reached human level [33, 30, 32, 27] and began to saturate. To further advance research, IJB-A dataset was introduced recently as a benchmark for unconstrained face recognition. Compared with the existing face datasets, IJB-A is quite challenging since it contains a wide range of appearance variations specially in the form of extreme head poses and variable image quality (see examples in Fig. 4). Since its release, the performances on IJB-A have gradually improved. The top performing methods on IJB-A employ learned feature representations from a large scale external database. For example, CNN features in combination with triplet loss embedding are used in [4, 29]. Chen *et al*. [3] use joint Bayseian metric learning along with CNN features. Five pose-specific CNN models are trained from facial data generated by 3D pose rendering in [1]. Features from a bilinear CNN architecture are used in [4]. The current top performing method [6] on IJB-A dataset uses a template adaptation strategy in combination with learnt features [27]. In order to compute a similarity score between two templates X and Y, it trains two binary classifiers $\mathcal{X}$ and $\mathcal{Y}$. Classifier $\mathcal{X}$ is trained using the media in X as positive class against a large negative media set. Classifier $\mathcal{Y}$ is trained in a similar fashion using the media in Y as the positive class. The similarity score between X and Y is then given by: $\frac{1}{2}\mathcal{X}(y) + \frac{1}{2}\mathcal{Y}(x)$, where $\mathcal{X}(y)$ is the similarity of template Y's media encoding (y) against classifier $\mathcal{X}$.

The IJB-A evaluation protocols are for template based face recognition, where both probe and gallery instances are represented with multiple visual items. Prior to the release of IJB-A dataset, image set classification based face recognition has been actively researched [40, 21, 2, 37, 41, 43, 9, 10, 11, 12]. Similar to a template, an image set is an unordered collection of multiple medias (such as mugshot images or video frames). While template (or image set) based classification provides many promises in the forms of multitude of data being readily available, it simultaneously poses modeling challenges emanating from the heterogeneity of such data in terms of both quality and content. A number of methods have been proposed in the literature to effectively model this information. For example, a template being represented on a non-linear manifold geometry (*e.g*. a point on the Grassmannian manifold [38] or Lie Group of Rieman-

nian manifold [37]) or by media combination (*e.g*. average pooling [8, 26]). In this paper, instead of representing all template medias by a single entity, we propose to keep it intact. The proposed scheme proves to be quite effective (evidenced by its superior performance in Sec. 5) since it avoids loss of any potential information contained in the template.

3. Joint Registration and Representation

Registration of a face to a canonical frontal view is quite crucial for the subsequent feature representation and classification steps. While the recently proposed data driven methods can automatically learn to represent faces, they resort to specially engineered techniques for registration. For example, DeepFace [33] warps a face to a canonical 3D model with the help of detected facial landmarks. In this paper, we propose to learn face registration jointly with the representation. For this purpose, we train a Convolutional Neural Network (CNN) which consists of two interconnected modules (Fig. 2). First, a registration module which learns a set of transformation parameters to optimally register a facial image. Second, a representation module which learns a distinctive feature encoding of the registered face image. The two modules are connected with the output of the registration module being input to the representation module. These modules are described next.

3.1. Registration Module

Registration of facial images typically involves cropping the most relevant facial region (with minimal background) and applying morphing operations on the cropped region to transform it to a canonical frontal view. This usually requires sophisticated facial pre-processing procedures (such as automatic landmark localization) which can be quite challenging, specially in the presence of extreme head poses. In this paper, we propose to adapt a dynamic learn-able mechanism, which automatically estimates a set of optimal parameters to spatially transform a given input face image. Our approach is CNNs based and deploys a Spatial Transformer Network [17] which has three parts: a *localization network* to regress a set of registration parameters. These parameters are then used by a *grid generator*, which outputs a sampling grid. Finally, a *sampler* which maps the input image onto the generated grid. The architecture of the localization network is shows in Fig. 3. Note that the first pooling layer implements mean pooling while the rest perform max operation. A pooling filter of 2×2 pixels is used in all layers. Each parameter layer is followed by a rectifier linear unit (ReLU) layer, except the final fully connected (FC) layer which regresses the transformation parameters.

For a given input image, the localization network outputs a set of six parameters of affine transformation, which are used to generate the sampling grid. The pixel values of the input image are then sampled onto the grid. This results in affine transformations (cropping, translation, rotation, scaling and skewing) of the input image. The registered face image then becomes an input to the subsequent representation module (described next).

3.2. Representation Module

In order to learn facial feature encoding, we employ VGG-16 [27]. It comprises of 8 convolutional and three fully connected layers, each of which is followed by one or more non-linearities (ReLU, pooling). With a relatively simple architecture, VGG-16 has shown superior performance on YTF and LFW benchmarks [27]. The complete network (with both the modules) is then trained using the publicly available face dataset by Parkhi *et al*. [27]. The dataset has 2.6 million face images of 2622 subjects. For training, the detected face regions (provided with the dataset) are loosely cropped. A cropped image contains full face region and may also have some background. The amount of background region is more in case of non-frontal and profile views. The registration module of the network is therefore deployed to only focus on the relevant facial region of interest and ignore any background. The subsequent representation module then learns a discriminative and distinctive feature encoding of the input face image. For an efficient training, we initialize the parameters of the representation module by VGG-Face model [27]. Parameters of the registration modules are initialized by seperately training it to output identity transformation parameters. After learning the parameters of the network, we consider the output of first fully connected layer of the representation module as feature encoding for an input image.

4. Template based Face Identification

A template is a set of images or video frames of the same subject. Face recognition from templates is relevant in scenarios where historical records of observations is readily available and should be leveraged to enhance systems performance. It becomes directly applicable in many real world commercial systems where multiple enrollments of a subject are simultaneously available. Examples include mugshots history of a criminal on the run in forensic search scenarios, lifetime enrollment images in national databases (passports, national identity cards and driver licenses) for access control systems, and multiple images of a person of interest in watch list scenarios (such as FBI's most wanted list). While multitude of heterogeneous data in a templates can be used to enhance face recognition performance, it simultaneously introduces many modeling challenges to make an effective use of this information. To leverage from this information, we propose to learn a discriminative model for each enrolled subject in the gallery and then deploy a

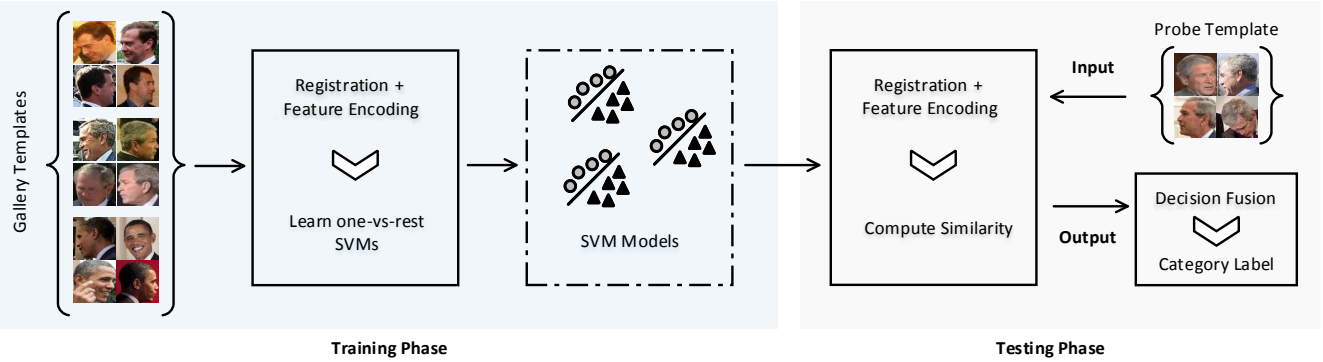


Figure 1: Block diagram of the proposed method. Class-specific discriminative models are learned after joint registration and feature encoding from a deep model during the training process. At test time, these models are used to compute similarity with the enrolled subjects and the individual decisions are combined to obtain a category label.

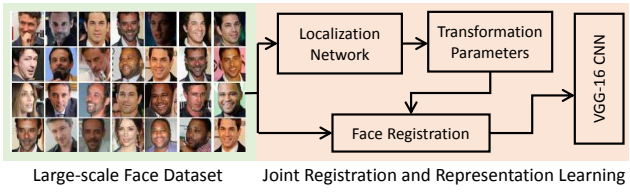


Figure 2: Joint face registration and representation.

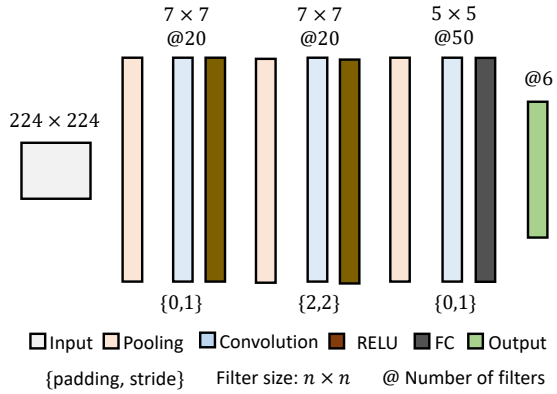


Figure 3: Localization network in the registration module.

score level fusion strategy for the probe templates. The details are given next.

4.1. Problem Description

For template based face identification, the gallery contains N templates $\{\mathcal{T}_1, \mathcal{T}_1, \mathcal{T}_1, \ldots \mathcal{T}_N\}$ corresponding to N enrolled subjects. Each template $\mathcal{T}_i = \{x_1, x_2, \ldots x_M\}$ has M medias (a media is an image or a video frame). Note that M is variable for each enrolled subject. At test time, we are given a query template $\mathcal{T}_q$, and the task is to find its best match with one of the enrolled gallery templates, or determine if $\mathcal{T}_q$ is not enrolled in the gallery.

4.2. Template Media Representation

Given a template $\mathcal{T}_i = \{x_m\} : m = 1 \cdots M$, we encode each media x_m by feed forwarding it through our trained Convolutional Neural Network model (as described in Sec. 3). The output of the first fully connected layer of the representation module is considered as the feature encoding for the template media. Given multiple template media encodings, there are different strategies proposed in the literature to effectively model them. Most of them find a suitable single entity representation for all template media. For example, all images and video frames in the template can be represented by a point on a geometric surface such as Grassmannian manifold [36], or Lie Group of Riemannian manifold [37]. The template media can also be represented by simply taking the mean of all media encodings [26, 8].

In this paper, instead of finding a single entity representation for heterogeneous template data, we propose to keep the media encodings intact. This helps avoid any loss of potential information contained in the template. In order to optimally use the multitude of data contained in the gallery templates, we propose to learn person specific models for each of the enrolled subjects in the gallery (details in Sec. 4.3). To optimally use the probe template data at classification, we employ a fusion strategy (details in Sec. 4.4). In our experimental evaluations (Sec. 5.2), we show that keeping the template media encodings intact is quite effective and results in significant performance boost.

4.3. Person-Specific Discriminative Models

For each of the enrolled subjects in the gallery, we learn a discriminative model. For this purpose, we train a simple one-vs-rest binary SVM classifier. Specifically, to learn the model parameters for a person, we consider feature en-

codings of all template medias for that person as the positive class, whereas the encodings of the remaining subjects are considered as the negative class. A binary SVM is then trained to learn a hyper-plane which optimally discriminates the two classes.

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_t \left(\max\left(0, 1-\ell_t \mathbf{w}^T\mathbf{x}_t\right)\right)^{\mathbf{2}}, \quad (1)$$

where $\ell_t = \{1, -1\}$. Following this procedure, we learn a set of model parameters $\{\mathbf{w}_i\} : i = 1 \dots N$ for N enrolled subjects in the gallery.

4.4. Query Template Classification

At classification, we are given a query template $\mathcal{T}_q = \{\mathbf{x}_m\} : m = 1 \cdots M$, where $\mathbf{x}_m$ is the encoding for mth media in the template. The task is to find $\mathcal{T}_q$'s best match with the enrolled gallery templates. Using our learnt person-specific models $\{\mathbf{w}_i\} : i = 1 \cdots N$, we can compute a decision value d_i^m for the mth template media to belong to ith enrolled subject. This is given by

$$d_i^m = \frac{1/\left(1+\exp^{-\mathbf{w}_i^T\mathbf{x}_m}\right)}{\sum_{i=\mathbf{1}}^{N} 1/\left(1+\exp^{-\mathbf{w}_i^T\mathbf{x}_m}\right)} \quad (2)$$

The above procedure gives us a set of decision values $\{d_i^m\} : m \in [1, M], i = [1, N]$. In order to combine these multiple decisions for all media in the query template, we explore two schemes. *First*, a simple mean of decision values approach, where given $\{d_i^m\}$, the predicted class label y_q of the query template $\mathcal{T}_q$ is determined by,

$$y_q = \arg\max_i \sum_m d_i^m. \quad (3)$$

Second, we employ a Bayesian approach inspired by the Bayesian Classifier Combination (BCC) model proposed in [19]. For each of the template media $\mathbf{x}_m$, we have a hidden true label $y_i \in [1, N]$ which matches it with an enrolled subject. We assume conditional independence between decisions d_i^m given the actual label y_i. Let us assume that y_i is generated by a multinomial distribution whose parameters are denoted by $\mathbf{p} : p(y_i = j|\mathbf{p}) = p_j$, where p_j represents the class probabilities (or proportions). Similarly, it can be assumed that decisions d_i^m for each media are generated by a multinomial distribution whose parameters are denoted by $\boldsymbol{\pi}_j^m : p(d_i^m = k|y_i = j) = \pi_{j,k}^m$. Note that $\boldsymbol{\pi}_j^m$ represents the rows of the confusion matrix $\boldsymbol{\pi}^m$ corresponding to each media representation. Therefore, the discriminative ability of each media representation is encoded in the Bayesian model.

The prior distributions for the parameters $\boldsymbol{\pi}_j^m$ and $\mathbf{p}$ are modeled by Dirichlet distributions with hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$p(\boldsymbol{\pi}_j^m|\boldsymbol{\alpha}_j^m) = \text{Dir}(\boldsymbol{\pi}_j^m; \boldsymbol{\alpha}_j^m) \quad (4)$$

$$p(\mathbf{p}|\boldsymbol{\beta}) = \text{Dir}(\boldsymbol{p}; \boldsymbol{\beta}) \quad (5)$$

Here, $\boldsymbol{\alpha}_j^m = [\alpha_{j,\mathbf{1}}^m \dots \alpha_{j,N}^m]$ and $\boldsymbol{\beta} = [\beta_{\mathbf{1}} \dots \beta_N]$. Further, we also define $\boldsymbol{\pi} = \{\boldsymbol{\pi}_j^m : j \in [1, N], m \in [1, M]\}$ and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_j^m : j \in [1, N], m \in [1, M]\}$. Then, we can define the joint posterior probability of the unobserved variables conditioned on the observed class decisions as:

$$p(\boldsymbol{y}, \mathbf{p}, \boldsymbol{\pi}|\boldsymbol{d}) \propto \prod_{i=\mathbf{1}}^{N} \left\{ p_{y_i} \prod_{m=\mathbf{1}}^{M} \pi_{y_i, d_i^m}^m \right\} p(\mathbf{p}|\boldsymbol{\beta}) p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \quad (6)$$

The original BCC model [19] utilizes Gibbs sampling for inference which is computationally expensive and slow in convergence. To achieve an efficient approximate inference, we use the Variational Bayesian (VB) formulation of Simpson *et al*. [31] which works similar to the Expectation Maximization (EM) algorithm. The VB approach analytically approximates posterior distribution $p(\boldsymbol{y}, \mathbf{p}, \boldsymbol{\pi}|\boldsymbol{d})$ (defined in Eq. 6) by a simpler and tractable distribution $q(\boldsymbol{y}, \mathbf{p}, \boldsymbol{\pi})$ which factorizes over its variables as follows:

$$q(\boldsymbol{y}, \mathbf{p}, \boldsymbol{\pi}) = q(\boldsymbol{y}) q(\mathbf{p}) q(\boldsymbol{\pi}) \quad (7)$$

where,

$$q(y_i = j) = \mathbb{E}_{\mathbf{y}}[y_i = j] = \rho_{i,j} / \sum_{k=\mathbf{1}}^{N} \rho_{i,k} \quad (8)$$

$$\text{s.t. } \rho_{i,j} = \exp(\mathbb{E}_{\mathbf{p}}[\ln p_j] + \sum_{m=\mathbf{1}}^{M} \mathbb{E}_{\boldsymbol{\pi}}[\ln \pi_{j,d_i^m}^m]) \quad (9)$$

$$q(\boldsymbol{p}) \propto \text{Dir}(\boldsymbol{p}; \boldsymbol{\beta}) \quad (10)$$

$$q(\boldsymbol{\pi}_j^m) \propto \text{Dir}(\boldsymbol{\pi}_j^m; \boldsymbol{\alpha}_j^m) \quad (11)$$

where the hyper-parameters are updated as follows:

$$\boldsymbol{\alpha}_j^m = \hat{\boldsymbol{\alpha}}_j^m + \left[\sum_{i=\mathbf{1}}^{N} \delta_{[\![d_i^m = k]\!]} \mathbb{E}_{\mathbf{y}}[y_i = j]\right]_{k=\mathbf{1}}^{N}$$

$$\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \left[\sum_{i=\mathbf{1}}^{N} \mathbb{E}_{\mathbf{y}}[y_i = k]\right]_{k=\mathbf{1}}^{N} \quad (12)$$

$\hat{\boldsymbol{\alpha}}_j^m, \hat{\boldsymbol{\beta}}$ denote the previous estimate of hyper-parameters. Using the current estimates of expectations in Eq. 8, we update the variational distribution in Eq. 7 (E-step). We then update the expectations in Eq. 8 as follows (M-step):

$$\mathbb{E}_{\mathbf{p}}[\ln p_j] = \frac{\Gamma'(\beta_j)}{\Gamma(\beta_j)} + \frac{\Gamma'(\sum_{k=\mathbf{1}}^{N} \beta_k)}{\Gamma(\sum_{k=\mathbf{1}}^{N} \beta_k)} \quad (13)$$

$$\mathbb{E}_{\boldsymbol{\pi}}[\ln \pi_{j,d_i^m}^m] = \frac{\Gamma'(\alpha_{j,d_i^m}^m)}{\Gamma(\alpha_{j,d_i^m}^m)} + \frac{\Gamma'(\sum_{k=\mathbf{1}}^{N} \alpha_{j,k}^m)}{\Gamma(\sum_{k=\mathbf{1}}^{N} \alpha_{j,k}^m)}, \quad (14)$$

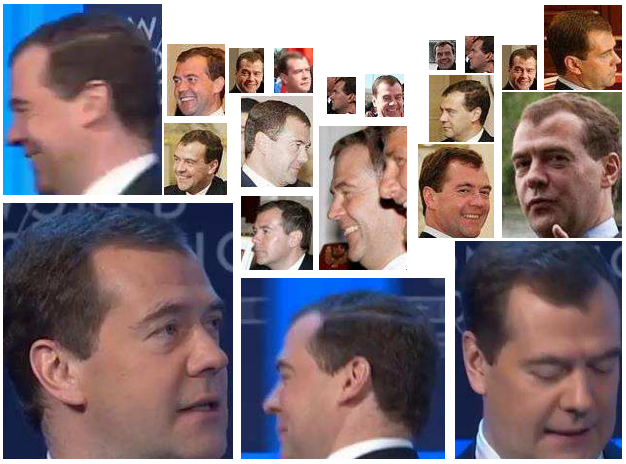


Figure 4: Sample Images of a person from IJB-A dataset. Note the extreme head poses and variations in image resolutions.

where $\Gamma(\cdot)$ is the standard gamma function used in the normalization constant of Dirichlet distributions. The VB algorithm for decision fusion works by iteratively updating the hidden output variables (actual labels y) and the model parameters $(\boldsymbol{\pi}, \mathbf{p})$.

5. Experiments

We extensively evaluate the performance of our proposed method on three datasets: IJB-A [22], YouTube Celebrities (YTC) [20] and COX [16]. For performance evaluation and comparison with existing state of the art, we use Cumulative Match Characteristics (CMC) and Decision Error Trade off (DET) curves. These metrics are defined in Sec. 5.2. Below, we first briefly describe the datasets used in our experiments.

5.1. Datasets

IJB-A dataset: contains 5712 images and 2085 videos of 500 subjects (from diverse geographic locations) captured in real life scenarios. While majority of other face recognition datasets contain either still images or video frames, IJB-A dataset contains both. The images and frames in the dataset exhibit diversity in terms of ethnicity, country of origin and head poses. The most challenging aspects of the dataset are the appearance variations caused by extreme head poses and variable image resolution. A few example images of a subject are shown in Fig. 4. In the presence of such extreme head rotations, automatic face detection fails quite often. The media in the dataset was therefore manually annotated to generate facial bounding boxes [22]. This avoids any frontal selection bias as a result of automated face detection failures in the presence of extreme head poses (e.g., in YTF and LFW datasets).

The IJB-A dataset is released with well-defined evaluation protocols. For template based face identification, 10 random training and testing splits are provided. Each split uses data of all 500 subjects with 333 subjects randomly sampled into the training set and the remaining 167 subjects form the testing set. The testing set contains probe and gallery templates. In order to make evaluation further challenging, 55 (randomly sampled) out of 167 subjects are removed from the gallery in the testing set. This corresponds to scenarios where probe subjects are not enrolled in the gallery. The probe templates of all 167 subjects are to be searched against the gallery templates of only 112 subjects.

YouTube celebrities [20] dataset contains 1910 videos of 47 celebrities downloaded from YouTube. Since the videos are acquired in real life situations, the resolution of the face images is very low and automatic face detection [34] fails for many videos. We therefore use tracking [28] to extract face regions from video frames. The extracted face region is then re-sized to 30×30 pixels. For template based face identification, we use five fold cross validation experimental protocol as in [14, 37]. Specifically, the complete dataset is divided into five equal splits with minimal overlap. Each split has nine templates (termed as image sets in [37, 14, 2]) per subject, three of which are used to form the gallery whereas the remaining six are the probe templates.

COX [16] dataset contains 4000 uncontrolled low resolution video sequences of 1000 subjects. In order to capture the videos, the subjects are asked to walk naturally inside a gymnasium without enforcing any constraints on their facial expression, lighting conditions and head poses. For our template based face identification experiments, we consider the frames of each video as a template and follow a leave-one-out strategy. Specifically, one template per subject is held-out as probe whereas the remaining form the gallery. For consistency, four runs of experiments are performed by swapping the probe and gallery templates.

5.2. Results

Evaluation Metrics: Face identification performance is commonly evaluated in terms of a Cumulative Match Characteristics (CMC) curve. A CMC curve plots identification rates corresponding to different ranks. A rank-k identification rate is defined as the percentage of probe searches whose gallery match is returned within the top-k matches. For scenarios where probes are not necessarily enrolled in the gallery, face identification performance is evaluated in terms of a Decision Error Trade-off (DET) curve, which plots False Negative Identification Rate (FNIR) vs False Positive Identification (FPIR) rate as a function of a similarity threshold for the top 20 candidates in the gallery. FPIR is the proportion of non-mate (not enrolled) probe searches returned above a similarity threshold. FNIR is the proportion

Table 1: Performance Evaluation on IJB-A dataset.

Methods	TPIR@FPIR=0.01	TPIR@FPIR=0.1	TPIR@Rank=1	TPIR@Rank=10
Bilinear-CNN [5]	14.2 ± 2.7	34.1 ± 3.2	58.8 ± 2.2	–
Face Search [35]	38.3 ± 6.3	61.3 ± 3.2	82.0 ± 2.4	–
Deep Multipose [1]	52.0	75.0	86.4	94.7
Triplet Similarity [3]	55.6 ± 6.5	75.4 ± 1.4	88.0 ± 1.5	97.4 ± 0.6
Joint Bayesian [29]	57.7 ± 9.4	79.0 ± 3.3	90.3 ± 1.2	97.7 ± 0.7
VGG-Face [6, 27]	46.1 ± 7.7	67.0 ± 3.1	91.3 ± 1.1	98.1 ± 0.5
Template Adaptation [6]	77.4 ± 4.9	88.2 ± 1.6	92.8 ± 1.0	98.6 ± 0.3
This Paper	88.6 ± 4.1	96.0 ± 1.0	96.4 ± 0.8	100.0 ± 0.0

.

of mate (enrolled) probe searches which are returned either below a similarity threshold or outside the top 20 ranks. For DET, we report True Positive Identification Rates (TPIR) at FPIR of 0.1 and 0.01, where TPIR$= 1-$FNIR.

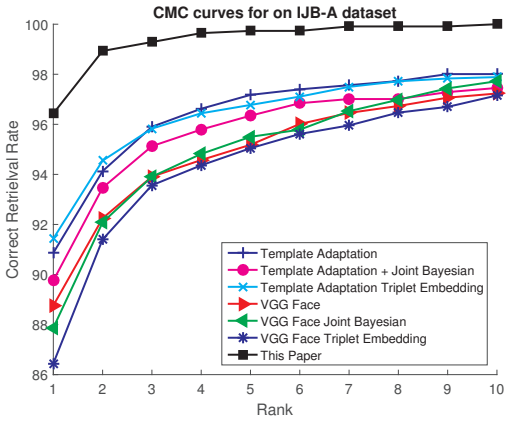


Figure 5: CMC curves on IJB-A dataset (best in colors).

Results on IJB-A Dataset: We compare the face identification performances on IJB-A benchmark in Table. 1. The results for the existing methods are reported from [6]. Due to a standard evaluation protocol on IJB-A dataset, the reported results are directly comparable. Our proposed method achieves average rank-1 and rank-10 identification rates of 96.4% and 100.0% respectively. For evaluations in the presence of non-mate probe searches, our method achieves average TPIR of 88.6% and 96.0% corresponding to FPIR of 0.01% and 0.1% respectively. Compared with the existing state of the art, the proposed method gains a relative performance boost of 3.9% (rank-1), 1.4% (rank-10), 8.8% (@FPIR=0.1) and 14.5% (@FPIR=0.01).

Results on YTC and COX Datasets: We further validate the efficacy of our proposed method on YTC and COX datasets. These datasets have been used in the literature for performance evaluation of image set classification methods. For the purpose of this paper, an image set can be considered as a template, as it contains multiple images or video frames. In Figure 6, we compare the performance of our method with a number of recently introduced image set classification methods. These include Mutual Subspace Method (MSM) [40], Discriminant Canonical Correlation Analysis (DCC) [21], the linear version of the Affine Hull-based Image Set Distance (AHISD) [2], Sparse Approximated Nearest Points (SANP) [14], Co-variance Discriminative Learning (CDL) [37], Regularized Nearest Points (RNP) [41], Set to Set Distance Metric Learning (SSDML) [43], Non-Linear Reconstruction Models (NLRM) [9] and Reverse Training (RT) [10]. For the compared methods, we use standard implementations provided by the respective authors. In order to encode facial images, we first use the original features proposed in the respective papers. We also evaluate the existing methods with our proposed features. The experimental results summarized in Figure. 6 show that our proposed method significantly outperforms the current state of the art by achieving average rank-1 identification rates of 90.1% and 83.6% on YTC and COX datasets respectively.

5.3. Discussion

We believe two major aspects of the proposed method contribute to its achieved superior performance. *First*, its strong feature representation capability. The proposed method learns to automatically register raw facial images while simultaneously finding a distinctive feature representation. Below, we show the effectiveness of the proposed features by evaluating them with existing methods. *Second*, its capability to synthesize multitude of information in the template media with proposed decision level fusion scheme. We further elaborate these aspects next.

Facial Feature Encoding: In order to demonstrate the effectiveness of our proposed learnt features, we evaluate them in conjunction with the existing image set classification methods in the literature. Specifically, instead of using the original features proposed in their respective papers, we use the facial features extracted by our method. By keeping the rest of the pipeline for the compared image set classi-

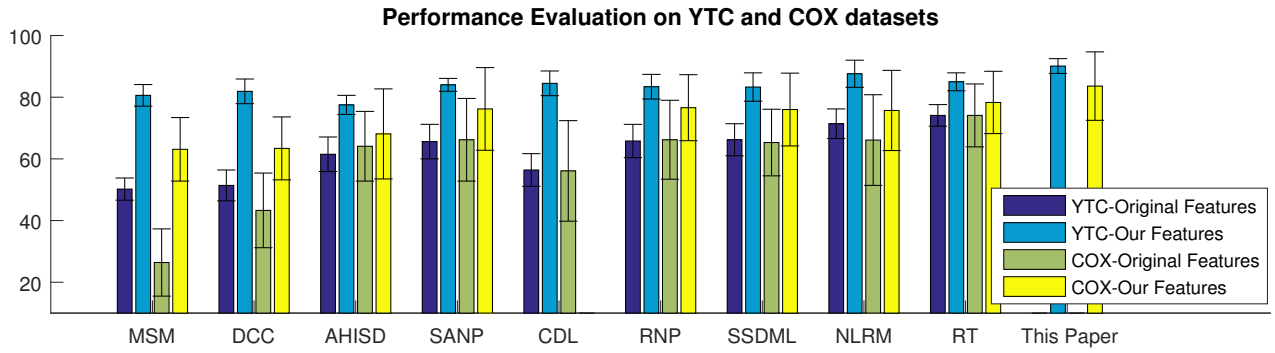
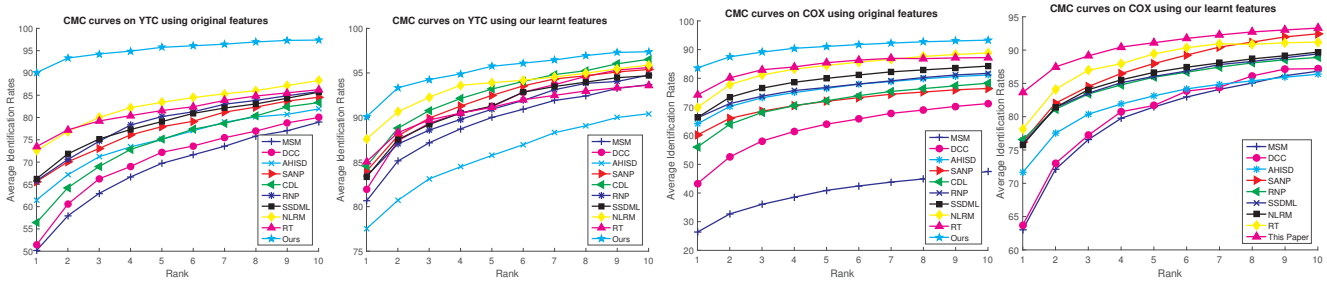


Figure 6: Rank-1 identification rates of different image set classification methods on YTC and COX datasets. Due to high memory requirements, CDL could not be evaluated on COX dataset with learnt features. Figure best seen in colors.



(a) YTC: Original Features (b) YTC: Learnt Features (c) COX: Original Features (d) COX: Learnt Features

Figure 7: CMC curves for different methods on YTC and COX datasets using their original features (a) & (c), and our learnt features (b) & (d). Figure best seen in colors.

fication methods fixed, our experimental results in Fig. 6 suggest that the performance of all methods significantly improves in combination with our proposed features. Note that due to large memory requirements, we were unable to evaluate CDL using learnt features on the COX dataset with available computational resources. CMC curves on the YTC and COX datasets in Figure 7 demonstrate that a consistent performance boost is achieved across all ranks.

Fusion - Feature vs Decision Level: For template (or image set) based face identification, multitude of information is present in the form of heterogeneous template media. Effectively utilizing this information is quite crucial to the overall face identification performance. In the existing literature, different strategies have been devised to find a suitable representation for the template media. These include a template represented by a point on a manifold geometry [38, 37], representative exemplars (*e.g.* derived from affine or convex hull models [2]) or by simply pooled media encodings [26, 8]. The existing methods therefore combine the information from multiple template medias at feature (media) level. In this paper, we keep the template media intact and do not find any single entity representation. Instead, we propose to synthesize information from all template medias at decision level. Even with the simple mean of decision values approach, we achieve a rank-1 identification rate of 94.2 ± 0.9 on IJB-A dataset. The proposed scheme to fuse information at decision level, instead of feature level, therefore avoids any potential loss of information and yields superior performance.

6. Conclusion

Template based face identification is pertinent in many real-world applications where multiple images of a persons' face are concurrently available, such as security and surveillance systems, watch list scenarios and access control systems. We presented a simple yet effective strategy to handle multitude of template media information. Unlike existing methods, which combine this information at initial feature level, we employed a Bayesian approach to fuse it later at decision level. For registration of unconstrained face data with extreme head poses, we presented a data driven approach to jointly learn registration with representation in a single Convolution Neural Network. Effectiveness of the proposed schemes is demonstrated by their significantly superior performance on challenging unconstrained face identification benchmarks.

References

[1] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, P. Natarajan, et al. Face recognition using deep multi-pose representations. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[2] H. Cevikalp and B. Triggs. Face recognition based on image sets. In *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on*, pages 2567–2573. IEEE, 2010.

[3] J.-C. Chen, V. M. Patel, and R. Chellappa. Unconstrained face verification using deep cnn features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[4] J.-C. Chen, R. Ranjan, A. Kumar, C.-H. Chen, V. M. Patel, and R. Chellappa. An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 118–126, 2015.

[5] A. R. Chowdhury, T.-Y. Lin, S. Maji, and E. Learned-Miller. One-to-many face recognition with bilinear cnns. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[6] N. Crosswhite, J. Byrne, O. M. Parkhi, C. Stauffer, Q. Cao, and A. Zisserman. Template adaptation for face verification and identification. *arXiv preprint arXiv:1603.03958*, 2016.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[8] T. Hassner, I. Masi, J. Kim, J. Choi, and S. Harel. Pooling faces: Template based face recognition with pooled face images. In *CVPR workshop*, pages 59–67. IEEE, 2016.

[9] M. Hayat, M. Bennamoun, and S. An. Learning non-linear reconstruction models for image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014.

[10] M. Hayat, M. Bennamoun, and S. An. Reverse training: An efficient approach for image set classification. In *European Conference on Computer Vision*, pages 784–799. Springer, 2014.

[11] M. Hayat, M. Bennamoun, and S. An. Deep reconstruction models for image set classification. *IEEE transactions on pattern analysis and machine intelligence*, 37(4):713–727, 2015.

[12] M. Hayat, S. H. Khan, and M. Bennamoun. Empowering simple binary classifiers for image set based face recognition. *International Journal of Computer Vision*, 2017.

[13] M. Hayat, S. H. Khan, M. Bennamoun, and S. An. A spatial layout and scale invariant feature representation for indoor scene classification. *IEEE Transactions on Image Processing*, 25(10):4829–4841, 2016.

[14] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(10):1992–2004, 2012.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report.

[16] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen. Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on COX-S2V dataset. In *Computer Vision–ACCV 2012*, pages 589–600. Springer, 2013.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015.

[18] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel. A discriminative representation of convolutional features for indoor scene recognition. *IEEE Transactions on Image Processing*, 25(7):3372–3383, 2016.

[19] H.-c. Kim and Z. Ghahramani. Bayesian classifier combination. In *International Conference on Artificial Intelligence and Statistics*, pages 619–627, 2012.

[20] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Computer Vision and Pattern Recognition (CVPR), 2008 IEEE Conference on*, pages 1–8. IEEE, 2008.

[21] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, 2007.

[22] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939. IEEE, 2015.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[24] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.

[25] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[26] E. Ortiz, A. Wright, and M. Shah. Face recognition in movie trailers via mean sequence sparse representation-based classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3531–3538, 2013.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, volume 1, page 6, 2015.

[28] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[29] S. Sankaranarayanan, A. Alavi, and R. Chellappa. Triplet similarity embedding for face verification. *arXiv preprint arXiv:1602.03418*, 2016.

[30] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[31] E. Simpson, S. Roberts, I. Psorakis, and A. Smith. Dynamic bayesian combination of multiple imperfect classifiers. In *Decision Making and Imperfection*, pages 1–35. Springer, 2013.

[32] Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900, 2015.

[33] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[34] P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[35] D. Wang, C. Otto, and A. K. Jain. Face search at scale. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[36] R. Wang and X. Chen. Manifold discriminant analysis. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 429–436. IEEE, 2009.

[37] R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503. IEEE, 2012.

[38] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[39] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011.

[40] O. Yamaguchi, K. Fukui, and K.-i. Maeda. Face recognition using temporal image sequence. In *Automatic Face and Gesture Recognition (FG), 1998 IEEE International Conference on*, pages 318–323. IEEE, 1998.

[41] M. Yang, P. Zhu, L. V. Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. pages 1–7, 2013.

[42] P. Yang, S. Shan, W. Gao, S. Z. Li, and D. Zhang. Face recognition using ada-boosted gabor features. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 356–361. IEEE, 2004.

[43] P. Zhu, L. Zhang, W. Zuo, and D. Zhang. From point to set: Extend the learning of distance metrics. In *International Conference on Computer Vision (ICCV), 2013 IEEE Conference on*. IEEE, 2013.